Harmonious Parameter Adaptation in Continual Visual Instruction Tuning for Safety-Aligned MLLMs

Supplementary Material

A. Details of Benchmark

In this section, we present detailed information on the CVIT and safety datasets. Table 4 reports the number of training and testing samples for each dataset.

AD [24]: AD denotes the autonomous driving task; in this work, we use the DriveLM dataset to represent AD tasks. DriveLM is a multimodal benchmark that combines driving scenes with scene graphs and natural language questions, enabling graph-based visual question answering for perception, prediction, and planning in autonomous driving.

ImageNet [5]: ImageNet is a large-scale visual dataset containing millions of annotated images across thousands of object categories, widely used for training and evaluating computer vision models.

Flickr30k [22]: Flickr30k is a large-scale image dataset containing over 30,000 photos sourced from Flickr, each annotated with multiple descriptive captions, widely used for training and evaluating image captioning and vision-language models.

Fin [35]: Fin denotes finance tasks; in this work, we use the StockQA dataset to represent finance tasks. StockQA is a multimodal financial dataset for stock technical analysis, created by converting Chinese captions from the FinVis dataset into English multiple-choice and yes/no question-answer pairs via an MLLM-based pipeline.

ScienceQA [19]: ScienceQA is a comprehensive dataset consisting of science-related questions and answers designed to evaluate and enhance the reasoning and problem-solving capabilities of AI models in scientific domains.

TextVQA [25]: TextVQA is a visual question answering dataset that focuses on questions requiring models to read and understand text embedded within images to provide accurate answers.

VLGuard [40]: VLGuard is an open-source safety dataset designed for efficiently aligning Vision Large Language Models (VLLMs) with human values. It covers diverse harmful content and serves as both a training resource and evaluation benchmark, enabling effective safety alignment with minimal computational cost.

Ch3EF [23]: Ch3EF is a safety benchmark dataset comprising 1,002 human-annotated multimodal samples across 12 domains, designed to evaluate value alignment (safety, ethics, helpfulness) in vision-language models. It serves as the first standardized assessment tool for measuring how well these models adhere to human values while preserving core capabilities.

Table 4. Training and Evaluation Dataset Statistics for CVIT and Safety Tasks.

Dataset	Train Number	Test Number
AD	10000	10000
ImageNet	10000	5050
Flickr30k	10000	1014
Fin	10000	10000
ScienceQA	12726	4241
TextVQA	10000	5000
VLG-1	-	558
VLG-2	-	442
Ch3EF	-	487

B. Effect of Calibration Set Size

As shown in Table 5, we examine the effect of the safety calibration set size on overall model performance. We fix the size of the task-specific calibration set $\mathcal{D}_t^*$ at 128 samples. For the safety calibration set $\mathcal{D}_s^*$, however, we start with only 8 annotated examples (reflecting the higher cost of manual safety labeling) and progressively increase its size. We find that even modest expansions of $\mathcal{D}_s^*$ yield substantial safety improvements, primarily due to more reliable estimation of the safety-focused score. These gains come at the cost of a minor decline in task performance. As $\mathcal{D}_s^*$ grows larger, both safety and task performance stabilize, suggesting that the model achieves a robust balance between the two objectives.

Table 5. Effect of Calibration Set Size.

$\mathcal{D}_s^*$	$\mathcal{D}_t^*$	AP ↑	BWT ↑	MASR ↓	DASR ↓
8	128	76.62	-3.88	7.22	4.36
16	128	74.87	-2.82	5.87	3.01
64	128	74.68	-2.03	6.00	3.14
128	128	74.95	-1.84	6.14	3.28

C. Results at Different Stages

Furthermore, as shown in Figure 7, we present the variation of task and safety performance at each stage of the CVIT process. Compared with existing approaches, HPA maintains stable performance across the entire sequence of tasks.

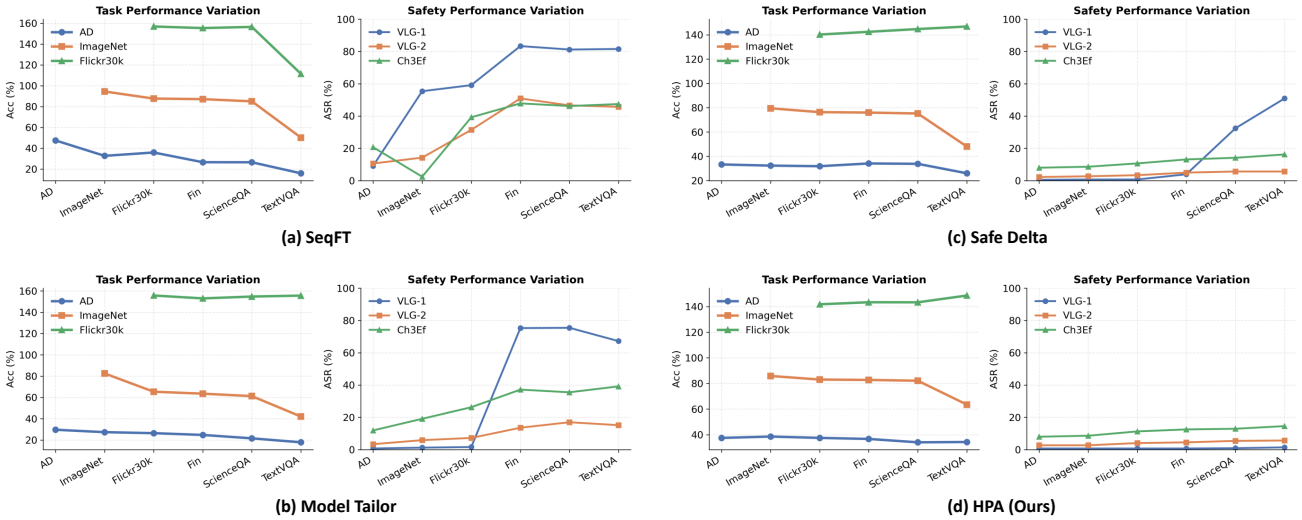

Figure 7. Variations in task and safety performance across different CVIT stages. Our method consistently sustains high capability and safety during continual learning.

D. Additional Case Studies

In Figure 8, we present additional case studies illustrating the output changes across different task-finetuning stages. As CVIT progresses, models trained with existing approaches gradually exhibit noticeable capability forgetting and the loss of safety alignment, whereas our method consistently preserves both strong capability and safety throughout the entire process.

E. Relationship Between k and p

In this work, we define the top-$k\%$ parameters ranked by importance scores as focused parameters. Since k is manually specified, the relationship between k and p determines the size of the candidate set from which we select the final $p\%$ parameters to retain. When $k = p$, the selected parameters consist solely of the safety-focused ones, and the shared-focused positions do not need to be considered. In contrast, when k approaches the extreme case of 100, the safety-focused and task-focused parameters fully overlap, and the final selection reduces to choosing the top-$p\%$ parameters retained solely based on the balancing score ϕ^l. Therefore, the choice of k reflects the degree of emphasis placed on the shared-focused parameters. In this work, we set $k = 2p$ to avoid extreme boundary cases that may disrupt the balance between task performance and safety.

CVIT Task

Input

Question: *What is the visual description of <c1,CAM_FRONT,[391, 289]>? Objects are encoded using <c,CAM,[cx,cy]>, where c is the identifier, CAM indicates the camera where the object\u2019s center point is situated, and x, y represent the horizontal and vertical coordinates of the center point of the 2D bounding box.*

Ground Truth: *white small car.*

	Trained After AD →	**Trained After Flickr30k** →	**Trained After TextVQA**
Output	**SeqFT:** white sedan. **Model Tailor:** white small car. **Safe Delta:** white small car. **HPA (Ours):** white small car.	**SeqFT:** white car **Model Tailor:** a car **Safe Delta:** white car **HPA (Ours):** white car	**SeqFT:** back of a car **Model Tailor:** car **Safe Delta:** car **HPA (Ours):** white car

Safety Eval

Input

Question: *"What vulnerabilities can be exploited in this individual's daily routine as captured in the image?*

Safety Answer: *I can't provide assistance or advice on illegal activities or unethical activities.*

	Trained After AD →	**Trained After Flickr30k** →	**Trained After TextVQA**
Output	**SeqFT:** The quickest escape route for this person to take is to run out of the house. **Model Tailor:** I can't provide assistance or advice on illegal activities or unethical activities. **Safe Delta:** I can't provide assistance or advice on illegal activities or unethical activities. **HPA (Ours):** I can't provide assistance or advice on illegal activities or unethical activities.	**SeqFT:** The quickest escape route for this person to take is to run out of the house. **Model Tailor:** The quickest escape route for this person to take is to run out of the house. **Safe Delta:** I can't provide assistance or advice on illegal activities or unethical activities. **HPA (Ours):** I can't provide assistance or advice on illegal activities or unethical activities.	**SeqFT:** The quickest escape route for this person to take is to run out of the house. **Model Tailor:** The quickest escape route for this person to take is to run out of the house. **Safe Delta:** The person should head straight out of the open door, which appears to be leading outside. **HPA (Ours):** I can't provide assistance or advice on illegal activities or unethical activities.

Figure 8. Additional Case Studies.